

（深圳）

实验报告

开课学期： 2024春季

课程名称： 统计机器学习

实验名称：构建决策树模型实现银行借贷预测

实验性质： 设计型

实验学时： 2 地点： T2102

学生班级： 计算机10班

学生学号： 220111028

学生姓名： 许辰涛

实验与创新实践教育中心制

2024年1月

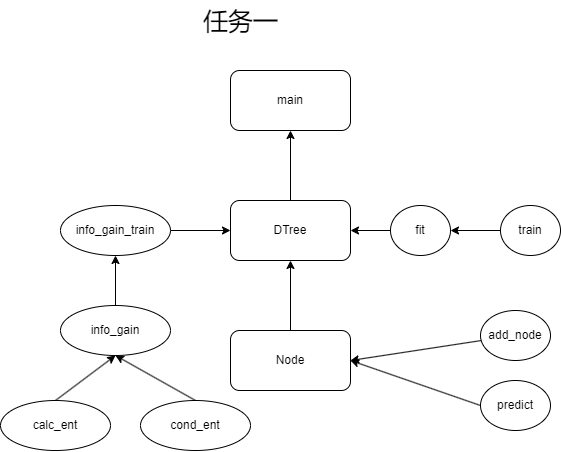
1. **实验环境**

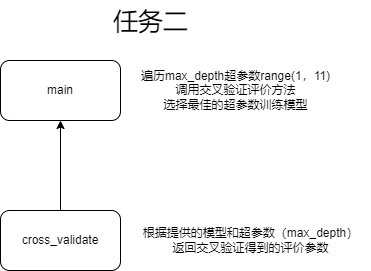
*请填写用到的操作系统和主要软件版本。*

操作系统：Windows11

Python：3.10.12

1. **实验过程**
2. *程序总体结构设计：main函数与各子函数之间的调用和返回关系，程序运行流程等，可以用文字、流程图、或图表的方式描述。*





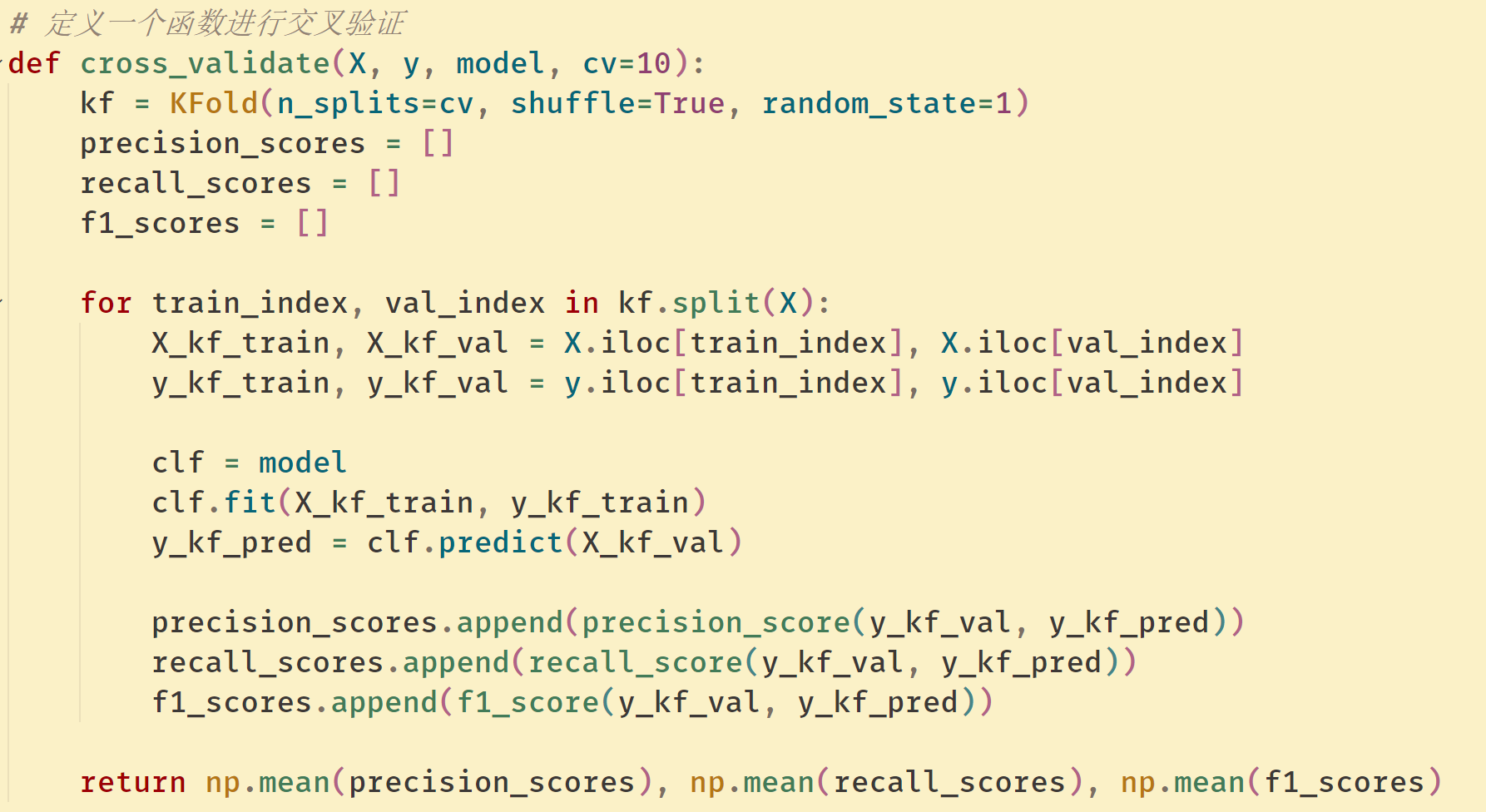
1. *数据集分割*



要点：

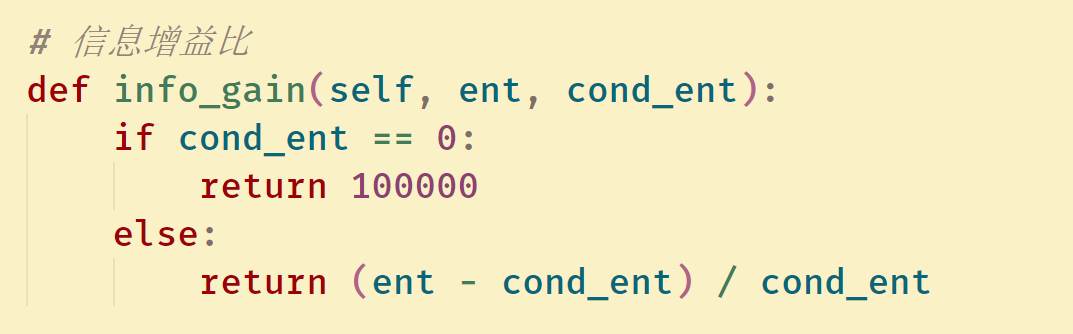
* revenue离散化
* 去除id列

K折交叉验证



1. *决策树算法：选用合适的决策树算法。*

任务1：

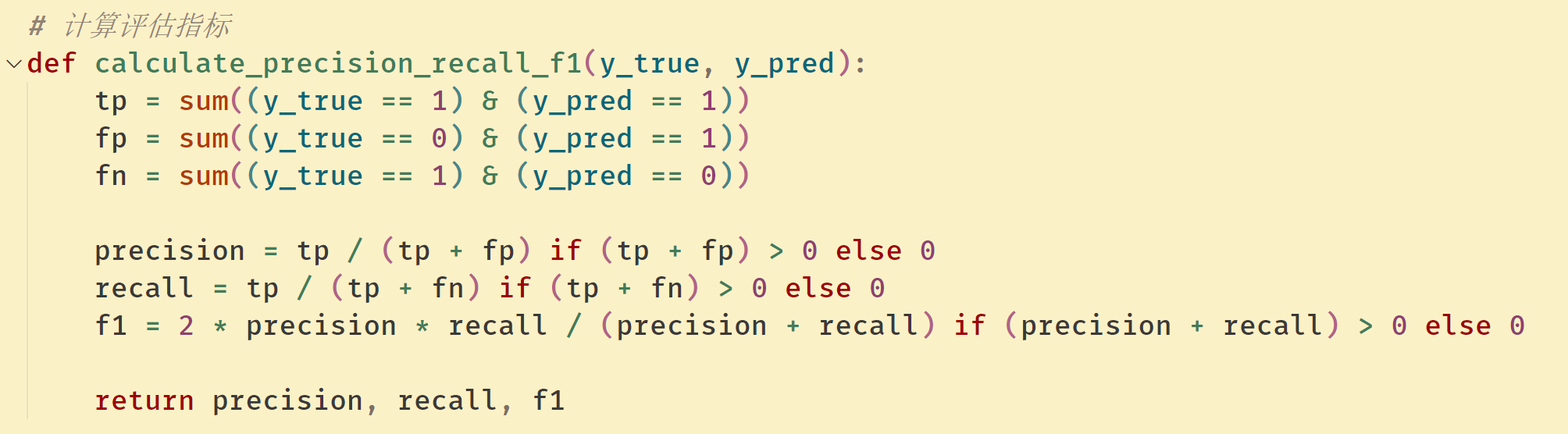


C4.5算法，挑选信息增益比最大的特征

任务2：

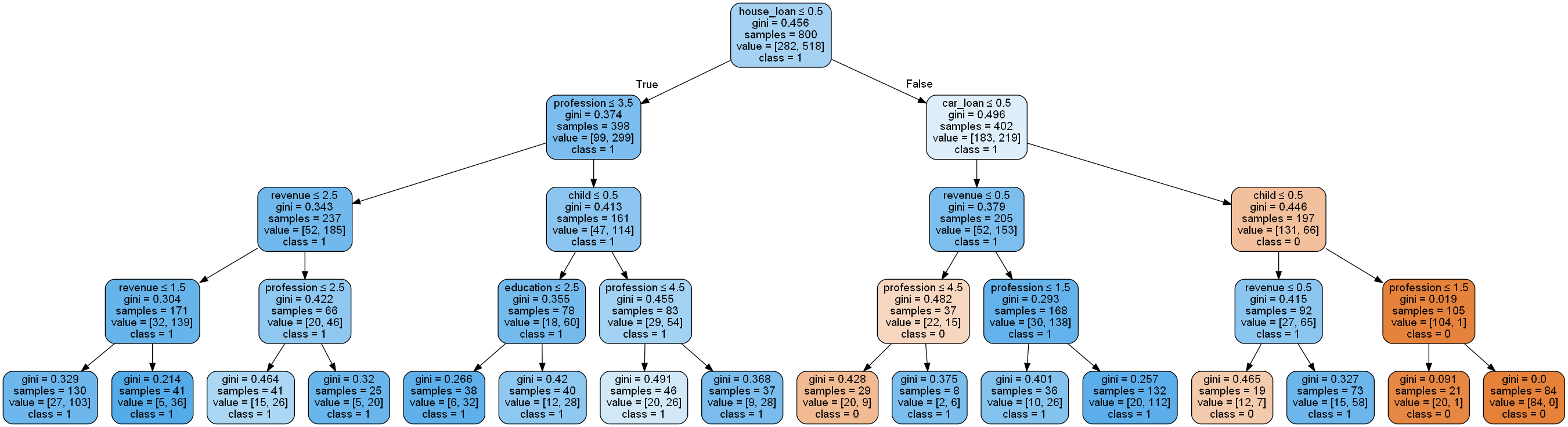
Scikit-learn中的DecisionTreeClassifier，使用优化后的CART算法

1. *评估模型：选用合适的评估指标。*



使用精确率、召回率得出的F1-score来评估模型

1. *总结分析：得到结论。*



借贷策略见决策树，在测试集上的指标为



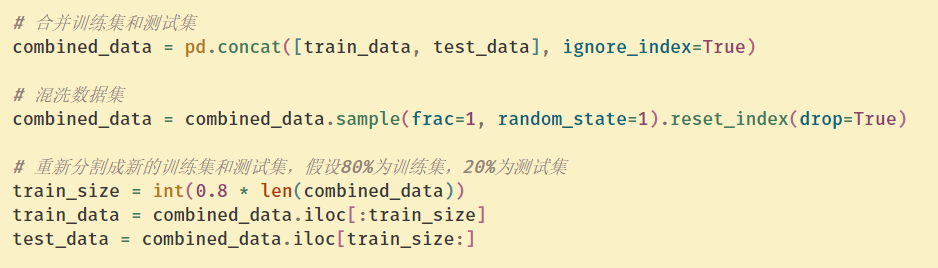
1. **收获与反思**

*请填写本次实验的收获，记录实验过程中出现的值得反思的问题及你的思考*

收获：熟悉了决策树算法，将理论与实践相结合，进一步熟悉了Python语法

值得反思的问题：任务二在训练集上的准确度并不高，在0.85左右，但是在测试机上却可以达到0.97，这使我产生了疑问，也不是欠拟合也不是过拟合

我的思考：考虑可能是训练集和测试集分布不均，可以尝试打乱后重新测试



在混合训练集测试集并打乱后得到了解决

